Comparative genome analysis of two Cryptosporidium parvum isolates with different host range

Published in final edited form as:
Infect Genet Evol. 2012 August ; 12(6): 1213–1221. doi:10.1016/j.meegid.2012.03.027.

Comparative genome analysis of two *Cryptosporidium parvum* isolates with different host range

Giovanni Widmer[a,*], **Yongsun Lee**[a], **Paul Hunt**[b], **Axel Martinelli**[c,*], **Max Tolkoff**[d,#], and **Kip Bodi**[e]

[a]Tufts Cummings School of Veterinary Medicine, Division of Infectious Diseases, North Grafton, Massachusetts 01536, USA

[b]Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh, UK

[c]Centro de Malária e Outras Doenças Tropicais, UEI Biologia Molecular, Lisbon, Portugal

[d]Tufts University Department of Computer Sciences, Medford, Massachusetts 02155, USA

[e]Tufts University Core Facility, Boston, Massachusetts 02111, USA

Abstract

Parasites of the genus *Cryptosporidium* infect the intestinal and gastric epithelium of different vertebrate species. Some of the many *Cryptosporidium* species described to date differ with respect to host range; whereas some species' host range appears to be narrow, others have been isolated from taxonomically unrelated vertebrates. To begin to investigate the genetic basis of *Cryptosporidium* host specificity, the genome of a *C. parvum* isolate belonging to a sub-specific group found exclusively in humans was sequenced and compared to the reference *C. parvum* genome representative of the zoonotic group. Over 12,000 single-nucleotide polymorphisms (SNPs), or 1.4 SNP per kilobase, were identified. The genome distribution of SNPs was highly heterogeneous, but non-synonymous and silent SNPs were similarly distributed. On many chromosomes, the most highly divergent regions were located near the ends. Genes in the most diverged regions were almost twice as large as the genome-wide average. Transporters, and ABC transporters in particular, were over-represented among these genes, as were proteins with predicted signal peptide. Possibly reflecting the presence of regulatory sequences, the distribution of intergenic SNPs differed according to the function of the downstream open reading frame. A 3-way comparison of the newly sequenced anthroponotic *C. parvum*, the reference zoonotic *C. parvum* and the human parasite *C. hominis* identified genetic loci where the anthroponotic *C. parvum* sequence is more similar to *C. hominis* than to the zoonotic *C. parvum* reference. Because *C. hominis* and anthroponotic *C. parvum* share a similar host range, this unexpected observation suggests that proteins encoded by these genes may influence the host range.

[*]Corresponding author: 200 Westboro Road, North Grafton, Massachusetts 01536, USA telephone 1 508 839 79 44; fax 1 508 839 7911; giovanni.widmer@tufts.edu.
[*]current address: University of Edinburgh, Edinburgh, UK
[#]current address: UCLA, Department of Biostatistics, Los Angeles, CA 90024

Keywords

Cryptosporidium parvum; *Cryptosporidium hominis*; genome; single-nucleotide polymorphism; ABC transporters

1. Introduction

Apicomplexan parasites of the genus *Cryptosporidium* multiply in the epithelium of the intestinal tract and are transmitted by the dispersal of oocysts in the environment and particularly in water. Many *Cryptosporidium* species are thought to have a narrow host range. This is reflected in the name of several species, such as *C. canis* (Fayer *et al.*, 2001), *C. felis* or *C. hominis* (Morgan-Ryan *et al.*, 2002). In contrast, other species appear to lack host specificity as they are naturally found in a wide range of host species or can be transmitted experimentally among taxonomically unrelated host species. Notable examples in the latter category are *C. parvum,* a zoonotic parasite of humans and livestock, *C. meleagridis*, which infects mammals and birds (Akiyoshi *et al.*, 2003) and the newly described species *C. ubiquitum* (Fayer *et al.*, 2010). In contrast to our expanding knowledge of the complexity of the genus *Cryptosporidium*, little progress has been made in identifying genetic loci controlling phenotypic traits (Okhuysen and Chappell, 2002). The advent of methods to rapidly re-sequence entire genomes may offer an alternative approach to the identification of genes controlling phenotypes in these parasites. This approach is enabled by the availability of the complete sequence of a *C. parvum* reference isolate (Abrahamsen *et al.*, 2004), the partially complete sequence of the *C. hominis* genome (Xu *et al.*, 2004) and the more recently sequenced *C. muris* genome.

Because *C. parvum* and *C. hominis* do not appear to recombine in nature, genome-wide association studies to find genetic determinants of host range are not feasible. A major polymorphic trait in *C. parvum* is host range. It has been known for some time that certain *C. parvum* genotypes are only found in humans (Mallon *et al.*, 2003). These observations have raised the possibility that the species *C. parvum* comprises two or more subgroups transmitted among different host species. “Anthroponotic *C. parvum*” is restricted to humans and is characterized by the presence of a distinct group of alleles of the sporozoite surface glycoprotein gene *GP60* (Cevallos et al., 2000; Strong et al., 2000) located on chromosome 6 (gene ID cgd6_1080). The alleles found in these isolates are assigned to the IIc genotype and have a characteristically short repeat of serine residues not found in *C. parvum* isolated from livestock (Sulaiman et al., 2005; Widmer, 2009). The restricted host range of *C. parvum* IIc contrasts with the zoonotic transmission of most other *C. parvum* genotypes. In nature, *C. parvum* isolates bearing the IIc *GP60* allele do not appear to recombine with zoonotic genotypes.

Here, to identify possible determinants of host range in *Cryptosporidium* parasites, the genome of the anthroponotic *C. parvum* isolate TU114 (*GP60* IIc (Widmer and Lee, 2010)) was sequenced and compared with that of the reference genome originating from the zoonotic *C. parvum* isolate IOWA. The SNP frequency across chromosomes reveals the presence of highly diverged loci. We postulate that some of these genes and intergenic regions may influence host range. Sequences of these loci in additional zoonotic and anthroponotic *C. parvum* isolates, as well as in taxonomically related *Cryptosporidium* species with different host range, will be required to assess the significance of these loci in defining the host range in this genus.

2. Material and Methods

2.1. Parasites

C. parvum isolate TU114 was isolated from a Ugandan child in 2003 and has since been maintained by serial propagation in immunosuppressed mice. This isolate was previously described in the context of a crossing experiment (Tanriverdi *et al.*, 2007). TU114 was genotyped with several polymorphic micro- and minisatellite markers (Tanriverdi and Widmer, 2006; Widmer and Lee, 2010). Its genotype is characteristic of a group of *C. parvum* isolates which are predominantly or exclusively transmitted among humans and carry the *GP60* (*GP40/15*) IIcA5G3b allele characterized by a short repeat of only 8 serine residues. Isolate TU114 has never been cloned. To generate DNA for sequencing, immunosuppressed CD-1 mice were orally inoculated with oocysts. When oocysts were first detected in the feces by the microscopic analysis of stained fecal smears (Ma and Soave, 1983), mice were transferred to cages with a wire bottom and feces collected daily. Oocysts were purified on step gradients of Histodenz (Widmer et al., 1998). To remove foreign DNA, oocysts were briefly suspended in a 10% solution of commercial bleach. Thereafter oocysts were subjected to three cycles of freezing and thawing, and DNA extracted with the HighPure DNA isolation kit (Roche Diagnostics, Indianapolis, IN).

2.2. Sequencing and mapping

Sequencing project 1: Genomic DNA from isolate TU114 was fragmented, ligated to Illumina adapters and the library loaded onto one lane of a flow cell. The library was sequenced in an Illumina Genome Analyzer$_{II}$ with 50-bp paired-end reads. A total of 22,125,448 reads were obtained, 21,483,810 (97.15%) of which mapped to the *C. parvum* reference genome (accession # NZ AAEE00000000). Hereafter, this sequencing project is referred to as “project 1”. Paired reads of 50 bp were aligned against the *C. parvum* IOWA reference genome using CLC Genomics Workbench version 4.03. Default alignment parameters were used except for the minimum and maximum distances for paired-end reads, which were set at 100 and 400 nucleotides (nt), respectively. The mean distance between paired reads was 200 nt. CLC SNP detection was run with default parameters except for the following settings: The minimum variant frequency was set at 80% and ploidy was set to 1. The “advanced” options were activated: the minimum paired coverage for calling SNPs was set to 4 reads, with a maximum coverage of 500 and a sufficient variant read count of 4.

Sequencing project 2: For sequencing project 2, a second Illumina library was generated from a different and unrelated sample of TU114 oocysts. In addition to TU114 oocysts, this sample contained an approximately equal number of oocysts of zoonotic isolate MD which were included because this library served as a control for an unrelated project. As is the case for the reference isolate IOWA, MD is a member of the zoonotic group of *C. parvum* and is genetically very similar to IOWA (Tanriverdi and Widmer, 2006; Widmer and Lee, 2010). Based on the alignment of about 20 kilobase (kb) of MD and IOWA sequence, we estimate that the two genomes differ by approximately 1 SNP/10 kb, or 1000 SNPs total. Thus, the number of SNPs introduced as a consequence of MD DNA being present was not expected to interfere with the identification of TU114 SNPs. This library was also sequenced with an Illumina Gene Analyzer$_{II}$. The library was sequenced single-end, in 40-nucleotide reads. A total of 19,546,814 reads were generated, 3,229,937 (16.5%) of which mapped to the *C. parvum* reference. The low percentage was due to the presence of contaminating DNA from the murine host and intestinal microflora remaining in the sample because, in this sample only, the surface sterilization step of the oocysts was inadvertently omitted. To identify the source of the contamination, the reads that did not map to *C. parvum* were assembled de novo using Velvet 1.0.17 (http://www.ebi.ac.uk/~zerbino/velvet/). This generated 2,510 contigs, of which 1,746 were longer than 1kb. A BioPerl script was used to successively

BLAST each contig against a downloaded copy of the non-redundant NCBI BLAST database. A total of 2,163 of the contigs mapped to "*Pseudomonas fluorescens* SBW25 complete genome", i.e., from a common intestinal bacterium. For the analysis of project 2 sequence reads the minimum variant frequency was set at 60% and the sufficient variant read count at 5. This lower threshold was intentionally chosen because this dataset served, as mentioned above, as a control for an unrelated experiment which requires the identification of low-frequency SNPs. Because of the low mean SNP coverage (mean = 9.7 reads), and the lower threshold, this dataset was used only to confirm SNP calls in project 1. SNPs identified in both sequences are referred to as "common". Those identified in one sequence only were excluded. As SNPs found in project 2 only were excluded from the analyses, any SNPs originating from MD were automatically excluded, as were low-confidence project 2 SNPs which would not have passed a more stringent threshold. The density of common SNPs (SNP/kb) was calculated across each chromosome in 50-SNP windows (1-SNP increments).

2.3. Other sequence analyses

Genetic distances between homologous *C. parvum* IOWA (reference), TU114 (anthroponotic *C. parvum*) and *C. hominis* (isolate TU502) sequences were calculated with Mega (Tamura *et al.*, 2007) using the Maximum Composite Likelihood substitution model (Tamura and Nei, 1993). A mismatch calculator (Schauer *et al.*, 2009) downloaded from www.famd.me.uk/agl/agl_sw.html was used to plot pairwise mismatch values in a 100-nucleotide or 33-amino acid sliding window. The window was shifted 5 positions at a time and the mismatch index *S* calculated using the "S - Gap" option. *S* is equal to the raw sum of mismatches (gaps ignored) summed over all positions in the window divided by the window size. S was converted to % such that the sum of *S* over the three pairwise comparisons equaled 100%.

To assess whether the distance between SNP and downstream open reading frame (ORF) differed among the four functional categories, distance values were compared using Kruskal-Wallis ANOVA on ranks. We compared the proportion of annotations among the highly diverged genes with the genome-wide proportion using a G-test of independence. Genes in functional categories were identified using the gene identification tools in cryptoDB.org (Puiu *et al.*, 2004).

3. Results

3.1. Sequencing results and quality control

Mapped to the IOWA reference genome, TU114 sequences from project 1 revealed 16,606 SNPs (Table 1). A total of 12,748 SNPs (76.8%) of these SNPs were also found in project 2 (Table 1). These common SNPs are the focus of our analyses. The remaining 23.2% SNPs were not found in sequencing project 2 and were excluded. Although some of the excluded project 1 SNPs may be genuine, having generated two independent Illumina sequences, we preferred a more conservative SNP-calling approach which excludes SNPs not identified in both projects. To assess the impact of excluding 23% of project 1 SNPs, we plotted SNP density for project 1 only, and for the common SNP dataset, and found only minor differences (Supplementary Fig. 1).

As low coverage could increase the number of sequence errors, we assessed whether certain genomic regions were under-sequenced. Sequence coverage for each common SNP in both sequencing projects was compared using linear regression. If a systematic bias in sequence coverage were present, we would expect to observe a correlation between coverage of homologous SNPs between projects. $\text{Coverage}_{\text{project 1}}$ was regressed on $\text{coverage}_{\text{project 2}}$ and the correlation coefficient R^2 calculated. For all chromosomes R^2 values were extremely

low, ranging from 0.0002 – 0.01 (Table 1), indicating that sequence depth was not correlated between sequencing projects and is random.

A preliminary Sanger sequencing survey of isolate TU114 (Tanriverdi et al., 2007) and alignment of single direction reads to 20,798 nt on chromosomes 1, 2 and 6 of *C. parvum* reference sequence (IOWA) called 26 SNPs (Supplementary Fig. 2). Of these, 13 were among the 14 'common' SNPs found in both Illumina sequencing projects. Six SNPs were identical to those found in one Illumina project only (3 in each) and 7 were absent from both projects. Repeat dideoxy sequencing confirmed the presence of one 'common' Illumina SNP that was not detected by original Sanger sequencing. Similarly, one SNP predicted by Sanger sequencing and not detected by Illumina sequencing was not confirmed (supplementary Fig. 2B). We conclude that, in these locations, the Illumina did not predict any false SNPs (false positives) and most likely missed few SNPs (false negatives). These data support the validity of the sequence generated on the Illumina platform.

3.2. Synonymous and non-synonymous SNPs

The percentage of non-synonymous SNPs between *C. parvum* IOWA and TU114 was found to be homogenous across chromosomes, ranging from 28% – 32% of all SNPs, with a genome-wide frequency of 30% (Table 1). Estimated with the method of Lipman (Nei and Gojobori, 1986) and the fact that 75.3% of the *C. parvum* genome is annotated as coding (Abrahamsen et al., 2004), we calculated that 60.0% of all nucleotide positions in the genome are non-synonymous. Therefore, consistent with the expectation that many non-synonymous sites are under negative selection, non-synonymous SNPs occur at about half the frequency as would be expected if all mutations were neutral. A striking resemblance in the distribution of all SNPs and non-synonymous SNPs was observed. Chromosomal scans of SNPs/kb generated similar curves if all SNPs or only non-synonymous SNPs were considered (supplementary Fig. 3).

3.3. Highly diverged regions

To identify highly diverged chromosomal regions, the frequency of SNPs (in sliding windows containing 50 SNPs) across the eight chromosomes was plotted (Fig. 1). This analysis enabled us to identify genes with a high frequency of SNPs. Using a threshold of 5 SNP/kb, which is 3.6-fold the genome-wide mean of 1.4 SNP/kb, we identified 27 regions with the highest SNP frequencies. These had a mean of 13.8 SNPs/kb (Table 2). The mean project 1 coverage of the SNPs located within these peaks was 29.6 reads, which is less than the average genome-wide coverage of 41.8 reads (Table 1). This suggests that extensive polymorphism in these regions may have interfered with the mapping of (highly divergent) reads to the IOWA reference genome. Consistent with this view, the average project 1 coverage of the *GP900* sporozoite surface glycoprotein gene (cgd7_4020), which encodes exceptionally long and polymorphic stretches of threonine repeats (Barnes *et al.*, 1998), was only 27.9. Moreover, a majority of SNPs in this gene were not called in project 2, because coverage was too low for the SNP calling criteria to be met. We also examined the possibility that the presence of low-complexity regions may have created peaks of artificially high SNP density. We identified low complexity regions using the Genome Browser in the cryptoDB database, which identifies such regions using the DUST algorithm (Morgulis *et al.*, 2006). Only one highly polymorphic region on chromosome 1 (reference position 111,916) was found spanning a low-complexity sequence. Because this sequence is only 17 nt long (reference position 111,895–111,912), whereas the peak extends over more than 28,000 nt, ambiguous mapping of sequence reads to this low-complexity sequence is unlikely to have artificially created a peak of this magnitude. We also investigated whether any high-SNP region was present in multiple similar copies across the genome, which could also have negatively affected the mapping of reads, but found no such cases. We conclude

from these analyses that highly divergent regions are not artifacts created by low complexity regions nor are they related to the presence of paralogous sequences.

Of the 27 regions displaying a SNP density exceeding 5 SNPs/kb, 22 (82%) peaked in annotated genes. The remaining 5 peaks were located outside annotated coding sequences, possibly indicating the presence of highly divergent intergenic regions, genes which are not annotated or transcripts which do not encode proteins. Two of these divergent and putatively intergenic regions were examined in more detail. According to EST data, the intergenic peak located at position 111,751 on chromosome 1 is transcribed in sporozoites (supplementary Fig. 4). This region may thus encode a gene which is currently not annotated. In contrast, the peak located at position 95,435 on chromosome 2 is centered on a short, transcriptionally silent intergenic region (not shown). This 434-nt region harbors 7 SNP, which converts to almost 16 SNP/kb, or more than 10-fold the genomewide average. However, the magnified SNP scans lack the resolution to pinpoint the exact location where SNP frequency peaks (not shown).

3.4. Properties of highly divergent genes

We investigated whether predicted structural and functional properties of the 22 genes located in the high-SNP regions deviated from the genome-wide expectation. Most striking was the large size of many of the highly divergent proteins. Their average relative molecular weight is 125,012, almost double the genome-wide average of 67,855 (Mann-Whitney Rank Sum test, p=0.002). On average there were more trans-membrane domains in the 22 highly divergent proteins as compared to the *C. parvum* proteome (1.22 vs. 0.79), but this distribution of trans-membrane domains was statistically not significantly different. Among the highly diverged genes 12 (54%) are annotated as “hypothetical”, a slightly higher, but statistically not significant, proportion than the 40.0% of genome-wide hypothetical annotation (Chi^2, p= 0.37). Among the 22 high-SNP genes, 3 (14%) are annotated as ATP binding cassette (ABC) transporter. This proportion is much higher than the 0.4% (n=16) genes genome-wide (p=0.0004). More than two-fold over-represented are genes encoding proteins with a signal peptide. Such genes represented 45% (n=10) of the 22 highly diverged genes, as compared to a genome-wide proportion of 21% (p=0.003).

3.5. Similarity to the human parasite *C. hominis*

Since isolates belonging to the IIc sub-group and *C. hominis* share a similar host range, we hypothesized that TU114 and *C. hominis* may share certain polymorphisms and that some or all of the highly divergent genes may influence the host range phenotype. Pairwise genetic distances between IOWA (zoonotic) *C. parvum, C. hominis* and TU114 were calculated to identify genes which are more similar between TU114 and *C. hominis* than between TU114 and zoonotic *C. parvum*. Overall, the Maximum Composite Likelihood distance values between the two *C. parvum* sequences (IOWA and MD) and *C. hominis* were typically about 10-fold larger than the distance between the *C. parvum* sequences and TU114. However, visual inspection of each sequence alignment indicated that in certain genes TU114 and *C. hominis* shared multiple SNPs. For instance, for the 675 5′-terminal nucleotides of gene cgd3_3370, TU114 is equally distant from IOWA *C. parvum* and *C. hominis* (Maximum Composite Likelihood distance = 0.021 and 0.022, respectively). To visually display the relationship between the three sequences, pairwise mismatch values *S* were calculated in 100-nucleotide and 33-amino acid windows. In the two examples shown in Fig. 2 (genes cgd3_3370 and cgd6_5260) there are several regions where there is greater similarity between *C. hominis* and TU114 sequences than between *C. parvum* and TU114. As expected, in a randomly chosen gene which lies in a conserved region of chromosome 3 (cdg3_2080; chromosome 3 reference position 545515–550854) no such pattern was observed.

The analysis of highly diverged genes was extended to a sample of six isolates originating from Ugandan children (Tumwine *et al.*, 2003). For this analysis we randomly selected among the highly diverged genes, gene cgd1_650 and gene cdg3_3370 (Table 2). The isolates were chosen from an archive of DNA samples of previously characterized isolates. Five of these isolates belong to the IIc subgroup of *C. parvum* but are genetically diverse based on their multi-locus genotype (Widmer and Lee, 2010). The sixth isolate (UG1862) does not cluster with the zoonotic or the anthroponotic *C. parvum* and carries a rare IIe allele at the *GP60* locus. In contrast to isolate TU114, these isolates originated directly from human patients and were not propagated in animals. Significantly, an alignment of the concatenated sequences totaling 409 nt obtained from these two genes revealed a close similarity between the sequence amplified from the six human isolates and *C. hominis* (Fig. 3). Even though the UG isolates are classified as *C. parvum,* their cgd1_650 and cgd3_3370 sequences were more similar to *C. hominis* than to isolate TU114. No difference was found between the *C. parvum* IOWA reference sequence and that from the zoonotic isolate MD.

3.6. Two-variant SNPs

Because *Cryptosporidium* parasites are haploid, the presence in isolate TU114 of SNPs with two variants is interpreted as an indication of intra-isolate heterogeneity, consistent with the fact that these isolates have not been cloned. A total of 224 SNPs, equivalent to 1.8% of all SNPs, have two variants in TU114. To assess whether this observation could be attributed to artifacts, as may occur in regions of low coverage, we compared the coverage of 2-variant SNPs with the mean coverage of all SNPs. This analysis showed that 2-variant SNPs had a slightly higher coverage (43.6 reads) than the genome-wide SNP coverage of 41.8 reads. The difference in coverage was statistically not significant (Mann-Whitney Rank Sum test, $p=0.33$), indicating that 2-variant SNPs are unlikely to be artefacts caused by low coverage.

A total of 85 2-variant SNPs located in 53 genes were identified as non-synonymous. Thus, in isolate TU114 1.4% of proteins are potentially polymorphic. The fraction of these genes with a product description “hypothetical” is 45% (n=24), which is similar to the proportion expected based on the number of the number of hypothetical *C. parvum* genes (n=1523, 40.0%). The proteins encoded by genes harboring 2-variant SNPs were significantly larger than expected based on the *C. parvum* proteome. As observed for the proteins encoded by the highly diverged genes, the median molecular weight of these polymorphic proteins is 149,035, more than twice the mean molecular weight of the remaining proteome (Mann-Whitney Rank Sum test, $p<0.001$). A total of 3 highly diverged genes feature on the list of the 53 genes harboring 2-variant SNPs. The probability of this outcome if the two properties were unrelated is only 3×10^{-3}, demonstrating that the association between these two properties is statistically significant.

3.7. Intergenic SNPs

The distance in nt separating each intergenic SNP from the next downstream ORF was calculated. Where SNPs were flanked by genes in head-to-tail orientation, i.e., encoded on the same strand, one distance value from the SNP to the next downstream translation initiation methionine codon was determined. If a SNP was located between two genes in opposite orientation, two distance values were computed if the genes were in head-to-head arrangement. No distance values were determined if the genes were oriented tail-to-tail. We analyzed these data to assess whether intergenic single-nucleotide mutations were associated with putative 5′ regulatory sequences. If this was the case, we reasoned that the distribution of SNP-ORF distances may vary for different functional categories, if functionally related genes are flanked by related regulatory sequences. The SNP-ORF distances belonging to different functional categories were grouped and compared. The categories ribosomal protein, kinase, ubiquitin and transmembrane were chosen because the number of genes is

relatively large (≥45), enabling a more robust analysis than for rare functions. As represented in Fig. 4, the frequency distribution of the SNP-ORF distance varied for different functional categories. The difference among functional categories is statistically significant (Kruskal-Wallis 1-way ANOVA on ranks, p=0.001). All functional categories were different to one another except kinase and ribosomal proteins.

4. Discussion

To identify candidate loci controlling host range, we compared the genome sequence of an anthroponotic isolate of *C. parvum* with the *C. parvum* reference genome which originates from a zoonotic isolate. To find SNPs we opted for a conservative approach and selected only SNPs found in two independent Illumina sequence datasets. The drawback of this strategy is the likely exclusion of genuine SNPs, particularly those called in Project 1 but not in project 2. As project 2 was sequenced using a single-end strategy, SNPs located in highly polymorphic regions may have been missed due to the difficulty of mapping short sequences to highly polymorphic regions. It is therefore possible that SNPs in such regions may be under-represented. This view is supported by gene cgd6_1080 which encodes a well-characterized sporozoite surface glycoprotein frequently termed *GP60* (Cevallos et al., 2000; Strong et al., 2000). In numerous surveys this gene has been shown to be very polymorphic. Yet in our analysis, cgd6_1080 scored a moderately elevated SNP density of 6.2/kb. This observation is consistent with the presence of extensive length polymorphism which may have interfered with the mapping of short Illumina reads. We also note that another surface glycoprotein gene encoding a highly polymorphic threonine-rich mucin (cgd7_4020) (Barnes et al., 1998) does not feature in Table 2. Because of this limitation, and the fact that deletions and insertions were not considered, highly repetitive sequences are likely to be underrepresented or missing from our analysis. Due to the limited resolution of the SNP density scans, highly diverged regions defined on the basis of a 5 SNPs/kb threshold may contain multiple genes. To eliminate any bias in calling highly diverged genes, only the locations where the SNP density was highest were considered. Genes flanking the peak, even if partially or completely located in a region where the SNP density exceeds the threshold, were not considered. This limitation is apparent in Supplementary Fig. 4, where several genes located upstream and downstream of a peak fall within a diverged region. A more stringently defined list of divergent genes will require the sequencing of additional genomes.

The only apparent phenotypic difference between the newly sequenced isolate TU114 and the reference is the host range. We postulate that diverged loci play a role in the adaptation of *Cryptosporidium* parasites to different host species and may therefore be under positive selection. The similarity between the divergent sequences in multiple *C. parvum* IIc (anthroponotic) isolates and the *C. hominis* sequence is consistent with this view as *C. hominis* and *C. parvum* IIc are human parasites. In light of the diversity in host range among *Cryptosporidium* species, identifying genes controlling this phenotype may shed light on host adaptation and speciation.

The theory of divergent adaptation predicts that this process eventually limits gene flow and may lead to reproductive incompatibility and speciation. It is unclear whether different *C. parvum* subgroups, such as *C. parvum* IIc, will eventually become reproductively isolated from the zoonotic subgroup. Natural *C. parvum* isolates bearing the IIc *GP60* allele do not appear to recombine with zoonotic genotypes in nature. However, the relatively small number of natural isolates genotyped using multi-locus methods, particularly from regions where IIc is common, is insufficient to rule out the existence of recombinants. In the laboratory, TU114 readily recombines with zoonotic *C. parvum* (Tanriverdi et al., 2007), suggesting that any speciation event has not led to reproductive incompatibility.

The identification of at least three genes (cgd1_650, cgd3_3370 and cgd6_5260) with high similarity between alleles from TU114 and *C. hominis* raises the possibility that their evolution is driven by the parasite's adaptation to different host species. This observation does not provide any clues on the evolutionary trajectory of these genes. The two possibilities that come to mind are (1) convergent evolution in a common host range, (2) a meiotic recombination event between zoonotic *C. parvum* and *C. hominis* which gave rise to *C. parvum* with restricted host range.

The host range of zoonotic *C. parvum* and anthroponotic *C. parvum* overlaps with that of *C. hominis* in humans. The evolutionary divergence between these taxa may represent different stages of sympatric speciation, where *C. parvum* and *C. hominis* are reproductively separated and the *C. parvum* genotypes are still able to recombine but do not appear to do so in nature. The host range of other *Cryptosporidium* species, such as the cattle parasite *C. bovis* (Fayer *et al.*, 2005), overlaps with that of zoonotic *C. parvum*. Sequencing of species with distinct and with similar host range is necessary to test the hypothesis that sympatric speciation is driven by the divergent evolution of a small number of loci like those identified in this study.

The over-representation of transporters among the highly divergent genes raises the possibility that these molecules play a previously unrecognized role in defining host range. In a parasite which is highly dependent on host cell metabolites, transporters play an important role in mediating the import of metabolites that the parasite is unable to synthesize (Striepen *et al.*, 2004). In contrast to the more extensively studied sporozoite surface glycoproteins (Wanyiri and Ward, 2006), less attention has been devoted to *Cryptosporidium* transporters (LaGier et al., 2002; Perkins et al., 1997). Similarly as sporozoite surface proteins mediate parasite invasion of the host cell and play an important role in evading the immune response, transporters play an important role during intracellular development and may therefore be involved in defining the host range. Evidence that ABC transporters are under selection was also found in an analysis of multiple *P. falciparum* genomes (Mu *et al.*, 2010). Our observations raise the possibility that genes expressed in intracellular stages may be more important than commonly assumed.

We based the identification of highly divergent loci on both synonymous and non-synonymous SNPs. The dN/dS ratio which is commonly used to identify genes under positive selection was not used because this metric is neither adequate for intra-species comparisons nor for comparing sequences from closely related genes (Kryazhimskiy and Plotkin, 2008; Wolf et al., 2009). As predicted by these authors, inspection of dN and dS rates (Yang and Nielsen, 2000) for 2580 pairs of orthologous *C. parvum* IOWA and TU114 genes showed that for several genes the dN/dS ratio was >1 because very few synonymous mutations were found. The non-random distribution of synonymous SNPs suggests that such SNPs are not necessarily neutral. Evidence that in the human genome synonymous SNPs may affect the phenotype has recently emerged from a meta-analysis of genome-wide association studies showing that synonymous SNPs were as likely to be associated with disease as non-synonymous SNPs (Chen *et al.*, 2010).

Intergenic SNPs were examined to evaluate whether the distribution of such sequence variation is affected by the functional group of the next downstream ORF. Such an effect would be expected if transcription of functionally related genes is driven by similar promoters and SNPs are less likely to occur within regulatory sequences. The significant difference in distribution of SNP-ORF distance between the five groups of genes we analyzed are consistent with this hypothesis and may indicate that at certain positions do not tolerate mutations due to the presence of regulatory motifs. Until additional *C. parvum*

genomes have been sequenced and analyzed, any interpretation of intergenic SNP distribution remains speculative.

In conclusion, the comparison of a newly sequenced genome from an anthroponotic *C. parvum* isolate with that of the zoonotic reference *C. parvum* has identified a small number of highly diverged genes. The three-way comparison of the orthologous sequences from *C. hominis,* zoonotic and anthroponotic *C. parvum* suggests that some of these genes may control host range. The significant over-representation of transporters among highly diverged genes suggests that the ability to establish an infection in a particular host species may depend in part on transporters controlling the exchange of metabolites between the host cell and intracellular developmental stages of the parasite. In the absence of reverse genetic methods to study *Cryptosporidium* parasites, additional *Cryptosporidium* genomes should be sequenced and compared to shed light on the mechanisms of host adaptation in this genus.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support from the NIAID (grant AI052781) is gratefully acknowledged. P.H acknowledges financial support from UK MRC and The Wellcome Trust. A.M. was supported by the Portuguese FCT. A preliminary sequence survey of isolate TU114 was performed in the laboratory of Michael Ferdig. Our thanks to Mark Blaxter and Urmi Trivedi from the University of Edinburgh GenePool for sequencing, to Joana Silva, Omar Harb and Steven Sullivan for critical comments, to Eric London for propagating TU114 and generating DNA, and to Lenore Cowen for facilitating student participation.

Abbreviations

SNP	single-nucleotide polymorphism
kb	kilobase
nt	nucleotide

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science. 2004; 304:441–445. [PubMed: 15044751]

Akiyoshi DE, Dilo J, Pearson C, Chapman S, Tumwine J, Tzipori S. Characterization of *Cryptosporidium* meleagridis of Human Origin Passaged through Different Host Species. Infect Immun. 2003; 71:1828–1832. [PubMed: 12654797]

Barnes DA, Bonnin A, Huang JX, Gousset L, Wu J, Gut J, Doyle P, Dubremetz JF, Ward H, Petersen C. A novel multi-domain mucin-like glycoprotein of *Cryptosporidium parvum* mediates invasion. Mol Biochem Parasitol. 1998; 96:93–110. [PubMed: 9851610]

Cevallos AM, Zhang X, Waldor MK, Jaison S, Zhou X, Tzipori S, Neutra MR, Ward HD. Molecular cloning and expression of a gene encoding *Cryptosporidium parvum* glycoproteins gp40 and gp15. Infect Immun. 2000; 68:4108–4116. [PubMed: 10858228]

Chen R, Davydov EV, Sirota M, Butte AJ. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. PLoS One. 2010; 5:e13574. [PubMed: 21042586]

Fayer R, Santin M, Macarisin D. *Cryptosporidium ubiquitum* n. sp. in animals and humans. Vet Parasitol. 2010; 172:23–32. [PubMed: 20537798]

Fayer R, Santin M, Xiao L. *Cryptosporidium bovis* n. sp. (Apicomplexa: Cryptosporidiidae) in cattle (Bos taurus). J Parasitol. 2005; 91:624–629. [PubMed: 16108557]

Fayer R, Trout JM, Xiao L, Morgan UM, Lai AA, Dubey JP. *Cryptosporidium canis* n. sp. from domestic dogs. J Parasitol. 2001; 87:1415–1422. [PubMed: 11780831]

Jukes, TH.; Cantor, CR. Evolution of Protein Molecules. In: Munro, HN., editor. Mammalian Protein Metabolism. Academic Press; New York: 1969.

Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. 2008; 4:e1000304. [PubMed: 19081788]

LaGier MJ, Keithly JS, Zhu G. Characterisation of a novel transporter from *Cryptosporidium parvum*. Int J Parasitol. 2002; 32:877–887. [PubMed: 12062559]

Ma P, Soave R. Three-step stool examination for cryptosporidiosis in 10 homosexual men with protracted watery diarrhea. J Infect Dis. 1983; 147:824–828. [PubMed: 6842020]

Mallon ME, MacLeod A, Wastling JM, Smith H, Tait A. Multilocus genotyping of *Cryptosporidium parvum* Type 2: population genetics and sub-structuring. Infect Genet Evol. 2003; 3:207–218. [PubMed: 14522184]

Morgan-Ryan UM, Fall A, Ward LA, Hijjawi N, Sulaiman I, Fayer R, Thompson RC, Olson M, Lal A, Xiao L. *Cryptosporidium hominis* n. sp. (Apicomplexa: Cryptosporidiidae) from Homo sapiens. J Eukaryot Microbiol. 2002; 49:433–440. [PubMed: 12503676]

Morgulis A, Gertz EM, Schaffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006; 22:134–141. [PubMed: 16287941]

Mu J, Myers RA, Jiang H, Liu S, Ricklefs S, Waisberg M, Chotivanich K, Wilairatana P, Krudsood S, White NJ, Udomsangpetch R, Cui L, Ho M, Ou F, Li H, Song J, Li G, Wang X, Seila S, Sokunthea S, Socheat D, Sturdevant DE, Porcella SF, Fairhurst RM, Wellems TE, Awadalla P, Su XZ. Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. Nat Genet. 2010; 42:268–271. [PubMed: 20101240]

Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986; 3:418–426. [PubMed: 3444411]

Okhuysen PC, Chappell CL. *Cryptosporidium* virulence determinants--are we there yet? Int J Parasitol. 2002; 32:517–525. [PubMed: 11943224]

Perkins ME, Volkman S, Wirth DF, Le Blancq SM. Characterization of an ATP-binding cassette transporter in *Cryptosporidium parvum*. Mol Biochem Parasitol. 1997; 87:117–122. [PubMed: 9233681]

Puiu D, Enomoto S, Buck GA, Abrahamsen MS, Kissinger JC. CryptoDB: the *Cryptosporidium* genome resource. Nucleic Acids Res. 2004; 32(Database issue):D329–331. [PubMed: 14681426]

Schauer SE, Schluter PM, Baskar R, Gheyselinck J, Bolanos A, Curtis MD, Grossniklaus U. Intronic regulatory elements determine the divergent expression patterns of AGAMOUS-LIKE6 subfamily members in Arabidopsis. Plant J. 2009; 59:987–1000. [PubMed: 19473325]

Sneath, PHAaSRR. Numerical Taxonomy. Freeman; San Francisco: 1973.

Striepen B, Pruijssers AJ, Huang J, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC. Gene transfer in the evolution of parasite nucleotide biosynthesis. Proc Natl Acad Sci U S A. 2004; 101:3154–3159. [PubMed: 14973196]

Strong WB, Gut J, Nelson RG. Cloning and sequence analysis of a highly polymorphic *Cryptosporidium parvum* gene encoding a 60-kilodalton glycoprotein and characterization of its 15- and 45-kilodalton zoite surface antigen products. Infect Immun. 2000; 68:4117–4134. [PubMed: 10858229]

Sulaiman IM, Hira PR, Zhou L, Al-Ali FM, Al-Shelahi FA, Shweiki HM, Iqbal J, Khalid N, Xiao L. Unique endemicity of cryptosporidiosis in children in Kuwait. J Clin Microbiol. 2005; 43:2805–2809. [PubMed: 15956401]

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007; 24:1596–1599. [PubMed: 17488738]

Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993; 10:512–526. [PubMed: 8336541]

Tanriverdi S, Blain JC, Deng B, Ferdig MT, Widmer G. Genetic crosses in the apicomplexan parasite *Cryptosporidium parvum* define recombination parameters. Mol Microbiol. 2007; 63:1432–1439. [PubMed: 17302818]

Tanriverdi S, Widmer G. Differential evolution of repetitive sequences in *Cryptosporidium parvum* and *Cryptosporidium hominis*. Infect Genet Evol. 2006; 6:113–122. [PubMed: 16503512]

Tumwine JK, Kekitiinwa A, Nabukeera N, Akiyoshi DE, Rich SM, Widmer G, Feng X, Tzipori S. *Cryptosporidium parvum* in children with diarrhea in Mulago Hospital, Kampala, Uganda. Am J Trop Med Hyg. 2003; 68:710–715. [PubMed: 12887032]

Wanyiri J, Ward H. Molecular basis of *Cryptosporidium*-host cell interactions: recent advances and future prospects. Future Microbiol. 2006; 1:201–208. [PubMed: 17661665]

Widmer G. Meta-analysis of a polymorphic surface glycoprotein of the parasitic protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. Epidemiol Infect. 2009; 137:1800–1808. [PubMed: 19527551]

Widmer G, Lee Y. Comparison of single- and multilocus genetic diversity in the protozoan parasites *Cryptosporidium parvum* and C. *hominis*. Appl Environ Microbiol. 2010; 76:6639–6644. [PubMed: 20709840]

Widmer G, Tzipori S, Fichtenbaum CJ, Griffiths JK. Genotypic and phenotypic characterization of *Cryptosporidium parvum* isolates from people with AIDS. J Infect Dis. 1998; 178:834–840. [PubMed: 9728554]

Wolf JBW, Kunstner A, Nam K, Jakobsson M, Ellegren H. Nonlinear Dynamics of Nonsynonymous (d(N)) and Synonymous (d(S)) Substitution Rates Affects Inference of Selection. Genome Biology and Evolution. 2009; 1:308–319. [PubMed: 20333200]

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA. The genome of *Cryptosporidium hominis*. Nature. 2004; 431:1107–1112. [PubMed: 15510150]

Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000; 17:32–43. [PubMed: 10666704]

Highlights

We sequenced the genome of an anthroponotic isolate of *Cryptosporidium parvum.*

The comparison of this sequence with that of the reference zoonotic *C. parvum* genome identified >12,000 single-nucleotide polymorphisms.

Genes located in highly diverged regions were identified and enriched gene functions among these genes identified.

The distribution of intergenic SNPs with respect to the downstream open reading frame was analyzed.

A 3-way comparison of *C. parvum* zoonotic, *C. parvum* anthroponotic and *C. hominis* identified polymorphisms which segregated with host range.

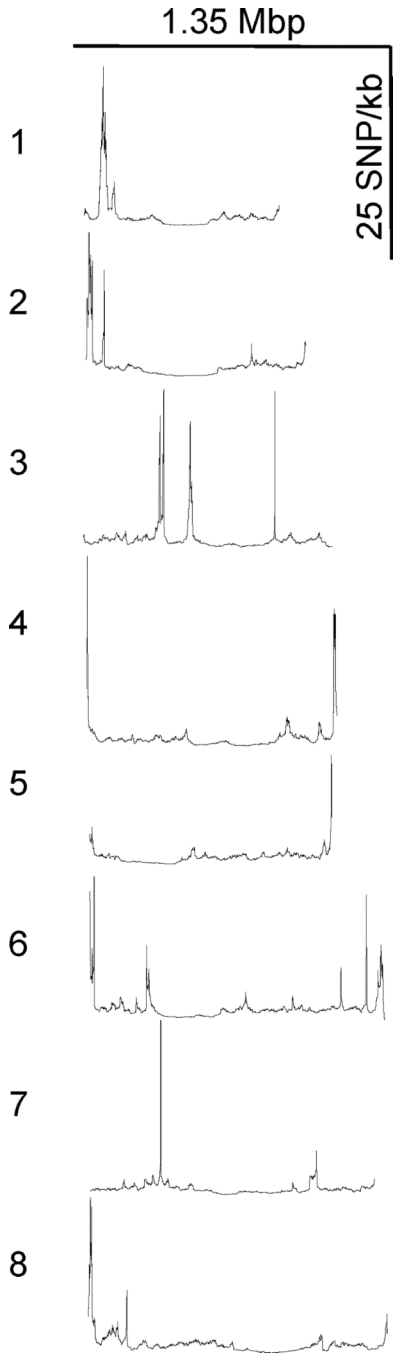

Fig. 1.

Genome-wide distribution of SNPs between *C. parvum* IOWA and TU114 genome. The scale for the x and y axes are shown uppermost. The eight chromosomes are drawn to scale. Note higher SNP frequency towards the ends of most chromosomes.

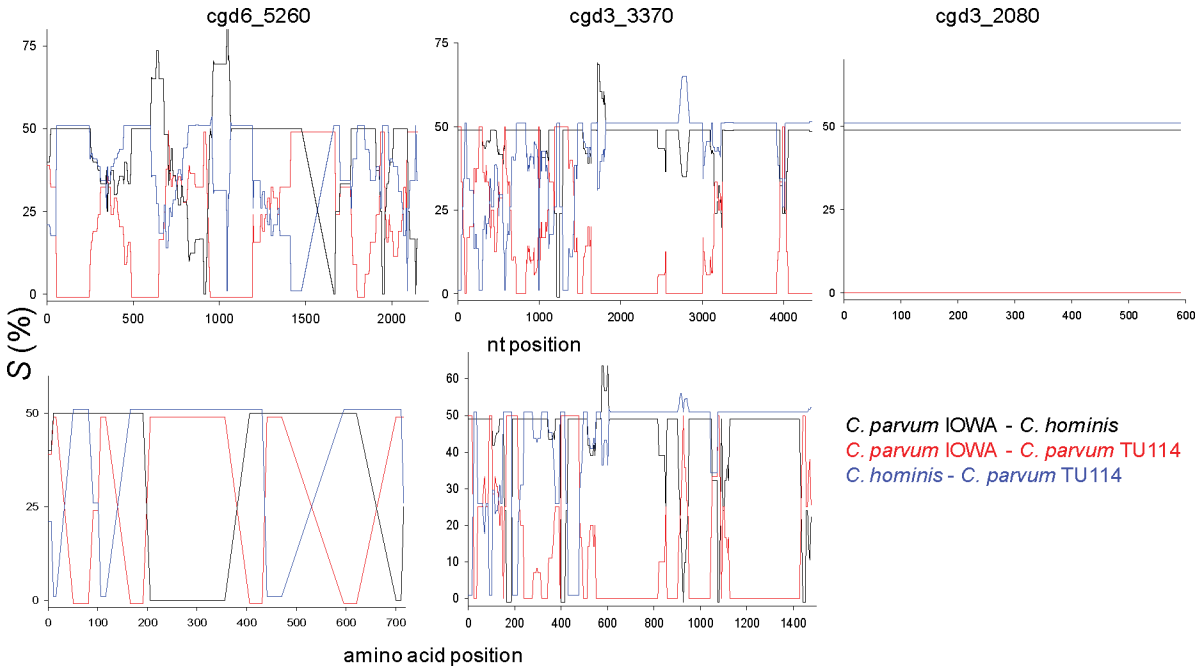

Fig. 2.

Three-way distance between *C. parvum* IOWA, *C. parvum* TU114 and *C. hominis*. Genetic distances in 100-nucleotide (upper graphss) and 33-amino acid (lower graphs) sliding windows were calculated as described by Schauer et al. (2009). Distance values *S* are expressed as percent of the sum of the three distance values. cgd3_2080 is located in a conserved region in the center of chromosome 3 (reference position 545515–550854) and is shown as a control. For cdg6_5260 and cgd3_3370 note the variability in the relative distance between the three isolates. Of particular interest is the presence of several regions where the relative distance between the two *C. parvum* sequences (IOWA - TU114, red) exceeds or is equal to the IOWA - *C. hominis* (black) or TU114 - *C. hominis* (blue) and where the blue line approaches zero. The blue and black lines were shifted in the vertical direction by ±1% to improve clarity.

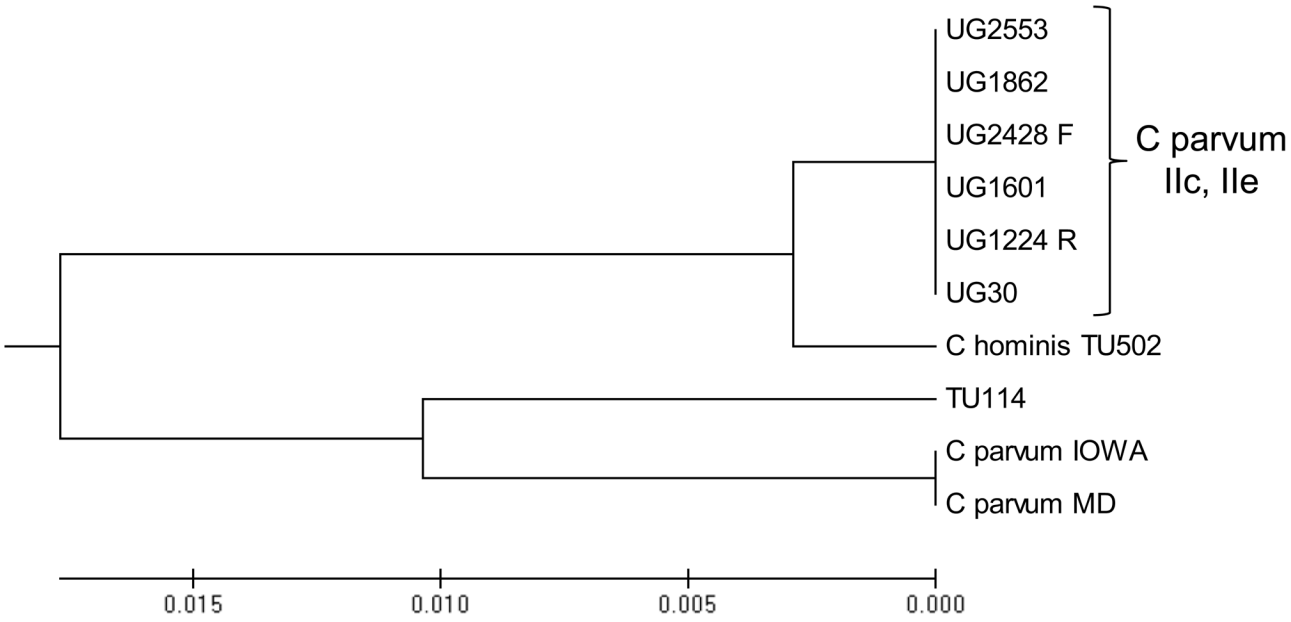

Fig. 3.

Phylogenetic analysis of highly diverged loci. The UPGMA method (Sneath, 1973) was used to display the similarity between portions of the highly diverged genes cgd1_650 and cgd3_3370 from nine *C. parvum* IIc isolates and from the *C. hominis* TU502 reference. The analysis is based on a 349-nt sequence comprising 14 polymorphic nucleotide positions. This sequence was obtained by concatenating a 241-nt fragment of gene cgd3_3370 and a 108-nt fragment of cgd1_650. Isolates designated with a UG code originate from human infections from Uganda and, except for UG1862, belong to the IIc sub-group. Isolate UG1862 has a unique multilocus genotype and its *GP60* allele belongs to the IIe group (Widmer and Lee, 2010). Except where indicated with a F (forward) and R (reverse), both strands were sequenced. The evolutionary distances are based on the Jukes-Cantor method (Jukes, 1969). The tree was drawn with MEGA4 (Tamura *et al.*, 2007).

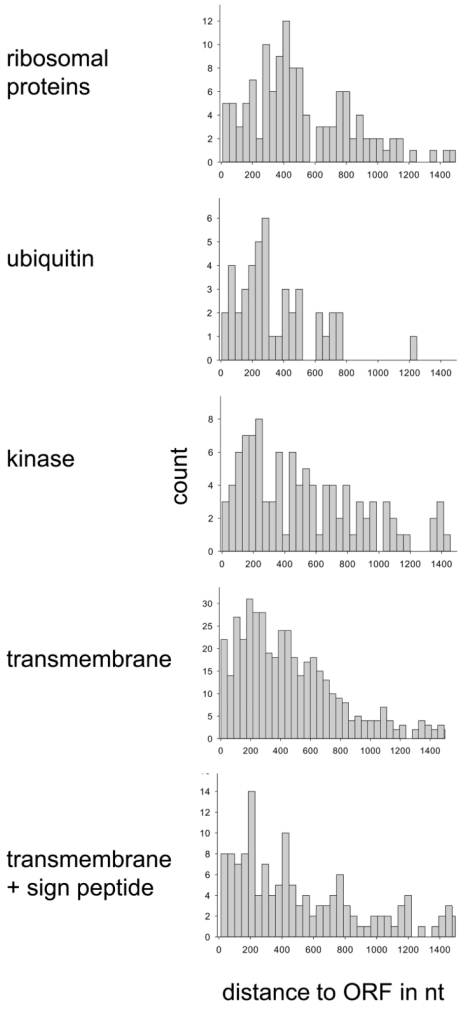

Fig. 4.

Distribution of intergenic TU114 SNPs. The histograms display the distance between each intergenic SNP and the nearest downstream translation initiation site for 4299 intergenic SNPs and five functional categories of the next downstream gene. Y axis, SNP counts; X axis, distance to ORF in nt.